American Journal of Medical Genetics 63:386–391 (1996)

Estimating Parental Relationship in Linkage Analysis of Recessive Traits

Chantal Mérette and Jurg Ott

Centre de Recherche Université Laval Robert-Giffard, Beauport, Quebec, Canada (C.M.); and Department of Psychiatry, Columbia University, and New York State Psychiatric Institute, New York, New York (J.O.)

In linkage analysis of recessive traits, parental relationship is important. For the case that it is unknown, the question is investigated as to whether estimating parental relationship and using the estimated relationship in linkage analysis is beneficial. Results show that estimating parental relationship can reliably be carried out on the basis of 50–100 genetic marker loci (analysis based on theory by Thompson [1975: Am J Hum Genet 39:173–188]). Misspecification of parental relationship leads to a loss of linkage informativeness, but not to false-positive evidence for linkage. An asymptotic bias in the recombination fraction estimate occurs when parents are unrelated and falsely taken to be related, but no such bias is seen when related parents are taken to be unrelated. Results from this investigation suggest that an estimated parental relationship may be used in linkage analysis as if it were the correct relationship, when evidence for the estimated relationship is supported by a likelihood ratio of at least 10:1 against the parents being unrelated. © 1996 Wiley-Liss, Inc.

KEY WORDS: linkage analysis, recessive disease, identity by descent

INTRODUCTION

For recessive diseases, it is well-known that inbred matings potentially provide much more information for linkage than noninbred matings [Smith, 1953; Lander and Botstein, 1986]. For this reason, families are sometimes collected in countries where it is relatively frequent for parents to be related. In many cases, however, the relationship between the parents is uncertain or unknown. Then, linkage analysis is typically carried out under the assumption that the parents are unrelated, which might unnecessarily reduce power if in fact the parents are related. This led us to consider a novel approach, i.e., to obtain an estimate for the parental relationship and use that estimated relationship in linkage analysis as if it were known.

Below, we investigate statistical properties of the proposed approach. Our investigation falls into two separate components: 1) determining the effect of misspecifying parental relationship in linkage analysis of recessive traits, and 2) estimating the relationship between two individuals. The latter question has been studied on the basis of genetic marker data [Thompson, 1975, 1986, 1991] and DNA fingerprinting [Chakraborty and Jin, 1993]. We will be concentrating on the marker-based approach.

Because an analytical investigation of these questions does not appear feasible, we employed computer simulation (Monte Carlo methods) to study properties of the proposed methods. We generate family data on the computer, i.e., under known conditions, and apply our methods to the generated family data. This allows us, for example, to tell with what probability our methods come to a correct or wrong conclusion. Such computer simulation methods belong to the standard repertoire of statistical geneticists.

MATERIALS AND METHODS

Misspecification of Parental Relationship

To study the effects of misspecifying parental relationship on linkage analysis of a recessive trait, we focused on two cases: 1) parents unrelated, and 2) parents as first cousins. Investigations were carried out on the two pedigree structures shown in Figure 1. PED1 refers to the individuals within the dotted line who form a nuclear family with 2 unrelated parents and 4 offspring, 3 affected and 1 unaffected, with marker information available on all 6 individuals. PED2 refers to the complete four-generation pedigree, where ancestors above the parents are deceased and thus unavailable for marker typing, whereas genetic marker information is available for the other individuals. For the PED2 pedigree structure, the parents of the 4 siblings are first cousins. For linkage analysis purposes, PED2 differs from PED1 only in the relationship of the parents of the 4 siblings.

Received for publication May 5, 1995; revision received October 27, 1995.

Address reprint requests to Dr. Jurg Ott, New York Psychiatric Institute, Unit 58, 722 West 168th St., New York, NY 10032.

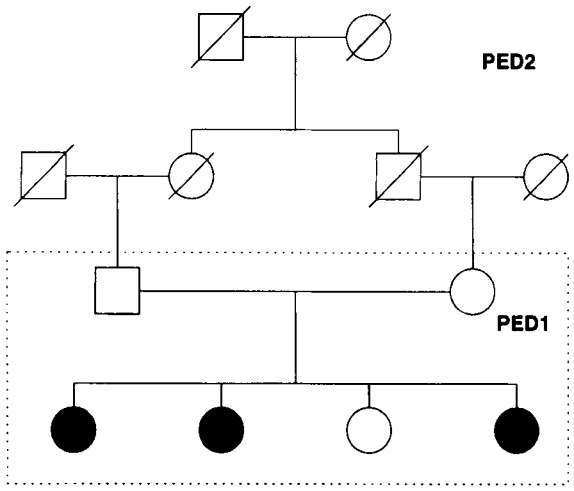

Fig. 1. Pedigree structures used for computer simulation. PED1 is limited to a nuclear family with unrelated parents, whereas PED2 incorporates grandparents and great grandparents, thus showing that the parents are cousins.

To study the effects of making correct or incorrect assumptions on the relationship of the 2 parents, we considered all four possible combinations between the two true situations, i.e., parents are in reality cousins or unrelated, and linkage analysis being carried out under the assumption that parents are cousins or unrelated. For each of the two true situations (PED1, parents unrelated; PED2, parents are cousins), for a single genetic marker locus with four equally frequent alleles (heterozygosity of 0.75), we generated marker genotypes by computer simulation using the SLINK computer program [Weeks et al., 1990]. Simulations were carried out for the following (true) recombination fractions: $r = 0$, 0.02, 0.04, 0.10, and 0.50. For $r < 0.50$, 500 replicates of the corresponding pedigree structures and marker genotypes were generated; for $r = 0.50$ (no linkage), 2,000 replicates were generated. Typically, each replicate contained only a single pedigree; for $r = 0.02$, replicates with n = 3, 5, and 10 pedigrees each were generated.

Each set of replicates was analyzed under each of two assumptions on the relationship between the 2 parents (unrelated, cousins) to learn of the effect of an incorrect assumption on parental relationship. For pedigree data generated under linkage ($r < 0.50$), in each replicate, lod scores were calculated at assumed (formal) recombination fractions, θ, ranging from 0–0.40 in steps of 0.02. At each value of θ, lod scores were averaged over all replicates, leading to an approximation of the expected lod score at that θ value. The maximum of these expected lod scores, the MELOD (a customary abbreviation for maximum of expected LOD score), was recorded as well as the θ value, $\tilde{\theta}$, at which it occurred. The difference, $\tilde{\theta} - r$, approximates the asymptotic bias in the estimate of the recombination fraction. Also, in each replicate, the observed maximum lod score, Z_{max}, over all θ values was recorded. Power and significance levels were estimated as the proportion of replicates in which Z_{max} exceeded some threshold, c. For power calculations ($r < 0.50$), we used $c = 3$, whereas significance levels ($r = 0.50$) were determined with respect to c = 0.5, 1.0, and 1.5.

Estimating Relationship

It is intuitively clear that 2 related individuals must have similar genotypes. Thus, estimation of an unknown relationship between 2 individuals (no relatives of either individual known) may be based on their genotypes at marker loci. Below, our approach is based on the theory developed by Thompson [1975, 1986, 1991], who derived the likelihood for the single-locus genotypes of 2 individuals of a given relationship.

For multiple unlinked marker loci, the total likelihood is simply the product over all single-locus likelihoods. For numbers of marker loci typically available in current human marker typing, we want to determine how reliably the relationship between 2 individuals can be estimated.

Consider a single genetic marker with a number of alleles, where its j-th allele, a_j, has population frequency, p_j. For a pair of individuals with a given relationship, assume, for example, genotype $a_u a_v$ for individual 1 and genotype $a_v a_w$ for individual 2. The likelihood for the observed genotypes is then simply the probability of occurrence, $P(a_u a_v, a_v a_w)$. This probability may be evaluated as follows: let q_i be the conditional probability of occurrence of the two genotypes, given that the individuals share i alleles identical by descent (IBD), $i = 0$, 1, or 2. Thompson [1991] provided a table of such conditional probabilities for all possible pairwise genotypes at a locus, where these probabilities depend only on IBD sharing and not on the relationship between 2 individuals. For example, when 2 individuals share no alleles IBD ($i = 0$), we have $q_0 = 4p_u p_v^2 p_w$. Further, $q_1 = p_u p_v p_w$ and $q_2 = 0$. Now, let k_i be the conditional probability that two individuals of a given relationship share i alleles IBD. Then, the joint probability of occurrence of the two genotypes for individuals of a given relationship is given by $k_0 q_0 + k_1 q_1 + k_2 q_2$, which is the conditional single-locus likelihood, given some relationship between the two individuals. For observed genotypes at multiple unlinked loci (we do not consider linked loci), the corresponding likelihoods are simply multiplied. The relationship between two individuals is now estimated by calculating the likelihood under a number of assumed relationships. The relationship with the highest likelihood is the estimated relationship.

In this approach, the relationship between two individuals is characterized by its associated IBD probabilities, k_i. Thus, relationships with identical IBD probabilities are indistinguishable, e.g., uncle-niece, grandparent-grandchild, and half-sibs [Thompson, 1986].

We investigated six different relationships between 2 individuals. Figure 2 shows a pedigree with 10 individuals containing examples of each of the relationships considered. Table I (based on Table I in Thompson [1991]) lists these relationships and their corresponding well-known IBD probabilities. Statistical proper-

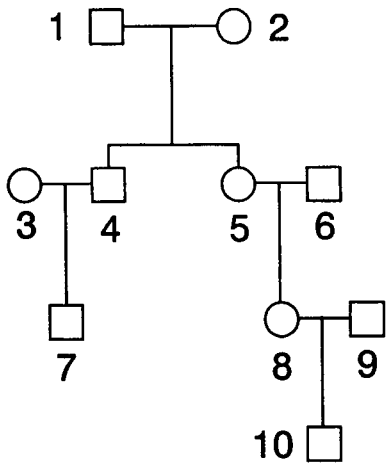

Fig. 2. Pedigree with ten members showing relationships listed in Table I.

ties of the maximum likelihood estimation procedure were determined by computer simulation as follows: using the SIMULATE computer program [Ott and Terwilliger, 1992], for each of the six (true) relationships of 2 individuals, we generated 500 replicates of genotypes at m unlinked marker loci, with each locus having 10 equally frequent alleles (heterozygosity of 0.90). For each set of 500 replicates, log likelihoods were calculated in six different ways, each time under the assumption of a different one of the six relationships considered here. These analyses were carried out for $m = 50$ and $m = 100$ markers. The latter number represents the maximum number of approximately unlinked marker loci over the human genome if we assume that, with a genome length of 40 Morgans, two markers are unlinked when they are at least 40 centimorgans apart.

RESULTS

Misspecifications of Parental Relationship

Parents are cousins. Table II shows results of the computer simulation for two true values of the recombination fraction, $r = 0$ and $r = 0.02$ (results for other values are referred to in the text). As Table II shows, the maximum expected lod score (MELOD) was always observed at the true recombination fraction, whether or not the analysis was performed under the correct familial relationship. This was the case for any of the recombination fractions, $r = 0$, 0.02, 0.04, 0.10 or 0.50 (results shown only for $r = 0$ and $r = 0.02$), i.e., ignoring in the analysis that parents are cousins and carrying out an analysis under the assumption of unrelated parents does not lead to an asymptotic bias. Not unexpectedly, however, there is a drop in the expected lod score when the analysis disregards an existing parental relationship. For example, at $r = 0$, the MELOD was 1.45 when the data were analyzed with the true parental relationship (first cousins). It decreased to 1.02 when in the analysis the parents were assumed unrelated, which represents a loss of 30% in informativeness for linkage. For the data simulated under $r = 0.02$, 0.04, and 0.10, the decrease in the MELOD was again close to 30% (results for $r > 0.02$ not shown).

TABLE I. IBD Probabilities, k_i, for Two Individuals With Selected Relationships*

Relationship	Examples	k_0	k_1	k_2
Parent-offspring	5, 8	0	1	0
Full sibs	4, 5	1/4	1/2	1/4
Uncle-niece[a]	4, 8	1/2	1/2	0
First cousins	7, 8	3/4	1/4	0
First cousins once removed	7, 10	7/8	1/8	0
Unrelated	3, 6	1	0	0

*Examples refer to numbered individuals shown in Figure 2.
[a]Also half-sibs, grandparent-grandchild.

TABLE II. MELOD and Recombination Fraction, $\tilde{\theta}$, at Which the Maximum Occurs, for True and Assumed Relationships of Parents (Alive) in Figure 1*

		Assumed relationship			
		Cousins		Unrelated	
True relationship	r	MELOD	$\tilde{\theta}$	MELOD	$\tilde{\theta}$
Cousins	0	1.45	0	1.02	0
Unrelated	0	0.55	0.08	0.95	0
Cousins	0.02	1.20	0.02	0.89	0.02
Unrelated	0.02	0.49	0.10	0.74	0.02

*r is the true recombination fraction.

Parents are unrelated. In contrast to the results discussed above, falsely assuming that the parents are cousins while in fact they are unrelated leads to an asymptotic overestimation of the recombination fraction. For example, with $r = 0$, we find $\tilde{\theta} = 0.08$. In addition, at $r = 0$, the MELOD drops from 0.95 (parents correctly treated as unrelated) to 0.55 (parents falsely assumed to be cousins), which represents a loss of 42%. For increased values of r, both asymptotic bias and drop in MELOD persist but tend to become less pronounced (results shown only for $r = 0$ and $r = 0.02$).

As Table II shows, analysis under the correct relationship is always better than under an incorrect relationship. In other words, analysis under an incorrect relationship does not tend to lead to inflated evidence for linkage; if anything, it may mask an existing linkage.

No linkage. Table III presents estimates of significance levels associated with critical maximum lod scores, $Z_{max} \geq c$, $c = 0.5$, 1.0, and 1.5 (2,000 replicates). For all cases shown, significance levels are smaller when assumed parental relationship is different from true relationship, i.e., assuming a wrong parental relationship for analysis purposes does not lead to false-positive evidence for linkage.

Power calculations. Table IV shows the results of power analyses for $n = 3$, 5, and 10 families of the structures given in Figure 1. When parents are first cousins, analysis under correct parental relationship yields power values of 0.78, 0.98, and 1.00, respectively. With an analysis assuming the parents to be unrelated even though they are in fact cousins, the corresponding power values are 0.54, 0.90, and 1.00. Thus, there is a

TABLE III. Significance Levels ($r = 0.50$) for True and Assumed Relationships of Parents (Alive) in Figure 1

True relationship	Assumed relationship: Cousins, lod score threshold 0.5	Cousins, 1.0	Cousins, 1.5	Unrelated, 0.5	Unrelated, 1.0	Unrelated, 1.5
Cousins	0.082	0.025	0.010	0.077	0.024	0
Unrelated	0.069	0.016	0.003	0.079	0.020	0

clear tendency for a decrease in power when parental relationship is not taken into account in the analysis. An analogous, but more pronounced, drop in power occurs when unrelated parents are falsely taken to be first cousins. These results confirm the tendencies seen above, i.e., that misspecification of parental relationship does not inflate evidence for linkage.

Estimating Relationship

As mentioned above, for a pair of individuals, marker data were generated under each of six relationships, R_t ("true relationship"). For each of the resulting 500 sets (replicates) of simulated data, the likelihood was calculated six times, each time assuming a different one of the six relationships ("stated relationship"), and the maximum likelihood estimate, $\hat{R}$, of the relationship was determined as that relationship with the highest associated likelihood. For each combination of true and stated relationships, the following two statistics were then calculated: S_1 was the probability that the stated relationship has the highest likelihood of all relationships investigated. This probability was approximated as the proportion of replicates in which the likelihood for the stated relationship was highest among all relationships considered. The second statistic, S_2, approximated the probability that the likelihood ratio, X = L (stated R)/L(R = unrelated), is $\geq$10. Thus, with 10 as a cutoff point for the likelihood ratio, S_2 may be viewed as the power in a significance test of the null hypothesis, R = unrelated, when individuals are related, or as the significance level when the true relationship is "unrelated," against the hypothesis, R = stated relationship.

As Table V shows, when 2 individuals are unrelated, S_2 is at most equal to 0.01, i.e., it is rare that some relationship is inferred (cutoff point, LR = 10) when in fact individuals are unrelated. Simulations were also carried out with a cutoff point of LR = 2.7 (corresponding to a difference in ln likelihood of 1; results not shown), but then falsely inferring a relationship occurred in up to 9% of the replicates.

When 2 individuals are siblings, their estimated relationship was always the true relationship. On the other hand, when a cutoff point of LR = 10 was applied for declaring a relationship accepted, no less than four relationships (sibs, uncle-niece, first cousins, and first cousins once removed) fulfilled this criterion. Thus, these relationships are difficult to distinguish for 2 siblings. This situation was the same for 50 or 100 marker loci.

For true first cousins, discriminating power among different relationships was better than for siblings. Only two relationships (first cousins, first cousins once removed) were detected with a power exceeding 50%. With 100 markers, power was estimated to be somewhat larger for first cousins once removed than for cousins, which, presumably, is due to a random fluctuation.

DISCUSSION

The main purpose of this investigation was to learn whether estimating an unknown parental relationship and using the estimated relationship in a linkage analysis of recessive traits might be beneficial. The answer depends on the true parental relationship. Assumption of a false parental relationship carries a penalty, in terms of both the maximum expected lod score (MELOD) (Table II), and power for detecting linkage (Table IV). The resulting loss of expected lod score is more serious when individuals are unrelated than when they are, for example, first cousins. Consequently, if estimated relationships are planned to be used in a linkage analysis, it is prudent to be conservative in declaring individuals to be related, for example, by applying a cutoff criterion of LR = 10, as in Table V, for the S_2 statistic.

Consider, for simplicity, the case that 2 parents are either first cousins, C, or unrelated, U. If they are C then, even with only 50 markers, the probability is 65% that this relationship will be detected (with the conservative criterion, LR $\geq$ 10). There is virtually no chance that they will be declared U (but they may be mistaken for having a relationship resembling C). In this case, estimating parental relationships is clearly beneficial. If the parents are U, there is little chance that they will mistakenly be called C. Thus, we conclude that it is generally useful to estimate the relationship between

TABLE IV. Power ($P[Z_{max} \geq 3]$) for Detecting Linkage*

True relationship	Assumed relationship: Cousins, Number of families 3	Cousins, 5	Cousins, 10	Unrelated, 3	Unrelated, 5	Unrelated, 10
Cousins	0.78	0.98	1.00	0.54	0.90	1.00
Unrelated	0.05	0.26	0.73	0.36	0.81	0.99

* $r = 0.02$ between disease and marker.

TABLE V. Resulting Statistics, S_1 and S_2, for Estimating Relationship Between Two Individuals*

		Stated relationship, R_s					
R_t	Statistic	Parent-offspring	Full sibs	Uncle-niece	First cousins	First cousins once removed	Unrelated
m = 50 markers							
Parent-offspring	S_1	1.00	0.00	0.00	0.00	0.00	0.00
	S_2	1.00	1.00	1.00	1.00	1.00	0.00
Full sibs	S_1	0.00	1.00	0.00	0.00	0.00	0.00
	S_2	0.00	1.00	1.00	1.00	1.00	0.00
Uncle-niece	S_1	0.00	0.00	0.88	0.12	0.00	0.00
	S_2	0.00	0.46	0.98	0.99	0.99	0.00
First cousins	S_1	0.00	0.00	0.11	0.61	0.25	0.03
	S_2	0.00	0.00	0.38	0.65	0.54	0.00
First cousins once removed	S_1	0.00	0.00	0.00	0.26	0.45	0.29
	S_2	0.00	0.00	0.06	0.19	0.13	0.00
Unrelated	S_1	0.00	0.00	0.00	0.02	0.22	0.76
	S_2	0.00	0.00	0.00	0.01	0.00	0.00
m = 100 markers							
Parent-offspring	S_1	1.00	0.00	0.00	0.00	0.00	0.00
	S_2	1.00	1.00	1.00	1.00	1.00	0.00
Full sibs	S_1	0.00	1.00	0.00	0.00	0.00	0.00
	S_2	0.00	1.00	1.00	1.00	1.00	0.00
Uncle-niece	S_1	0.00	0.00	0.95	0.05	0.00	0.00
	S_2	0.00	0.48	1.00	1.00	1.00	0.00
First cousins	S_1	0.00	0.00	0.05	0.75	0.19	0.01
	S_2	0.00	0.00	0.38	0.87	0.91	0.00
First cousins once removed	S_1	0.00	0.00	0.00	0.19	0.59	0.22
	S_2	0.00	0.00	0.03	0.26	0.35	0.00
Unrelated	S_1	0.00	0.00	0.00	0.00	0.17	0.83
	S_2	0.00	0.00	0.00	0.01	0.01	0.00

*R_t, true relationship; R_s, stated relationship; $\hat{R}$, estimated relationship; S_1, $P(\hat{R} = R_s)$; S_2, $P(L[R_s]/L[\text{unrelated}] \geq 10)$.

parents whenever it is in any way doubtful that they are unrelated. To localize the gene for a recessive disease on the human genome, we recommend the following steps, assuming that nuclear families are investigated:

1) With a number of unlinked markers, estimate the relationship between the parents.
2) For analysis purposes, if the estimated relationship has a likelihood at least 10 times that for unrelated parents, make up relatives of the parents such that they show the estimated relationship. If the likelihood ratio is <10, treat the parents as being unrelated.
3) Carry out linkage analyses using the modified family tree, i.e., with the estimated relationship between the parents.

Typically, model misspecifications lead to an asymptotic bias in the recombination fraction estimate [Ott, 1991]. As is seen above, when parents are cousins but in the analysis assumed to be unrelated, the recombination fraction estimate is still asymptotically unbiased. An explanation for this phenomenon has been given previously (section 10.6 in Ott [1991]): when parental phase probabilities are different from ½ but in the analysis assumed to be equal to ½, the resulting recombination fraction estimate was shown to be asymptotically unbiased.

For many populations, there are good records of the frequencies of various parental relationships. Consider, for example, the prior probability, p = P(C), that two parents are cousins. In our simplified situation,

$$1 - p = \text{P(U)}$$

is then the prior probability that they are unrelated. Such priors may be used to obtain more precise estimates of parental relationship. If L(C) = P(data | C) is the likelihood for the marker data given that 2 individuals are cousins, and L(U) is the likelihood given that they are unrelated, then the posterior probability that they are cousins is given by

$$P(\text{C} \mid \text{data}) = \frac{\text{pL(C)}}{[\text{pL(C)} + (1 - \text{p})\text{L(U)}]}.$$

This equation is easily extended to several relationships.

In the analysis of extended pedigrees, many errors in determining marker genotypes tend to be exposed as Mendelian incompatibilities. No such safeguards exist in the estimation of relationship between 2 individuals. Thus, marker errors should be kept to as low a level as possible. One might even consider building marker errors into the estimation procedure, but this is not pursued here. Similarly, errors in allele frequencies might influence the results of estimating parental relationships. However, as shown by Thompson [1975], such estimates are robust against changes in allele frequencies.

ACKNOWLEDGMENTS

We thank Dr. Linda Brzustowicz for approaching us with the question of how one might estimate relation-

ship among 2 individuals, which led us to find the literature already available on this subject and eventually prompted us to investigate properties of assuming an estimated parental relationship in linkage analysis.

This work was supported by grant HG00008 from the National Center for Human Genome Research, by the National Retinitis Pigmentosa Foundation, and by the Fonds de la Recherche en Santé de Québec, Canada.

REFERENCES

Chakraborty R, Jin L (1993): Determination of relatedness between individuals using DNA fingerprinting. Hum Biol 65:875–895.

Lander ES, Botstein D (1986): Mapping complex genetic traits in humans: New methods using a complete RFLP linkage map. Cold Spring Harbor Symp Quant Biol 51:49–62.

Ott J (1991): "Analysis of Human Genetic Linkage." Baltimore: Johns Hopkins University Press, pp 217–227.

Ott J, Terwilliger JD (1992): Assessing the evidence for linkage in psychiatric genetics. In Mendlewicz J, Hippius H (eds): "Genetic Research in Psychiatry." New York: Springer-Verlag, pp 245–249.

Smith CAB (1953): The detection of linkage in human genetics. J R Stat Soc 15:153–184.

Thompson EA (1975): The estimation of pairwise relationships. Ann Hum Genet 39:173–188.

Thompson EA (1986): "Pedigree Analysis in Human Genetics." Baltimore: Johns Hopkins University Press, pp 47–55.

Thompson EA (1991): Estimation of relationships from genetic data. In Rao CR, Chakraborty R (eds): "Handbook of Statistics," Vol 8. New York: Elsevier, pp 255–269.

Weeks DE, Ott J, Lathrop GM (1990): SLINK: A general simulation program for linkage analysis. Am J Hum Genet 47:204.